CLAIMS

1. A method for encoding characters appearing in an area of an image in order to generate a corresponding output string of character codes, the method comprising:

identifying one or more sequences of the character codes that are likely to be generated due a segmentation error in application of a pattern recognition process, and associating a respective extension character code with each of the sequences;

dividing the area of the image into segments such that each segment contains approximately one character;

applying the pattern recognition process to each of the segments in order to generate an input string of character codes, the input string comprising a respective character code for each of the segments;

locating at least one of the sequences of the character codes in the input string, and replacing the at least one of the sequences with the respective extension character code so as to generate a modified string; and

determining the output string by comparing the modified string to a directory of known strings.

2. The method according to claim 1, wherein the character codes that are generated by the pattern recognition process are selected from a predetermined set of eight-bit codes, and wherein associating the respective extension character code comprises assigning a respective eight-bit code that is not included in the predetermined set to replace each of the sequences.

3. The method according to claim 1, wherein applying the pattern recognition process comprises applying optical character recognition (OCR).

4. The method according to claim 1, wherein determining the output string comprises finding an approximate match between the modified string and one of the known strings, and outputting the one of the known strings.

5. The method according to claim 4, wherein finding the approximate match comprises computing respective edit distances between the modified string and a plurality of the known strings, and selecting the one of the known strings responsively to the respective edit distances.

6. The method according to claim 5, wherein computing the respective edit distances comprises determining respective costs of edit operations involving the extension character code, and applying the respective costs in computing the respective edit distances.

7. The method according to claim 6, wherein each of the one or more sequences of the character codes is generated due to incorrect segmentation of a respective original character having a respective original character code, and wherein determining the respective costs comprises assigning a cost of zero to a transformation of the respective extension character code associated with each

of the sequences to the respective original character code.

8. The method according to claim 4, wherein finding the approximate match comprises:

replacing each of the one or more sequences of the character codes in the known strings with the respective extension character code so as to create aliases that are respectively derived from the known strings;

adding the aliases to the directory; and

finding the approximate match between the modified string and one of the aliases, and

wherein outputting the one of the known strings comprises outputting the one of the known strings from which the one of the aliases is respectively derived.

9. Apparatus for encoding characters appearing in an area of an image in order to generate a corresponding output string of character codes, the apparatus comprising:

a memory, which is arranged to hold a directory of known strings; and

at least one processor, which is arranged to receive an identification of one or more sequences of the character codes that are likely to be generated due a segmentation error in application of a pattern recognition process, and to associate a respective extension character code with each of the sequences, and which is further arranged to divide the area of the image into segments such that each segment contains

approximately one character, to apply the pattern recognition process to each of the segments in order to generate an input string of character codes, the input string comprising a respective character code for each of the segments, to locate at least one of the sequences of the character codes in the input string, and to replace the at least one of the sequences with the respective extension character code so as to generate a modified string, and to determine the output string by comparing the modified string to the known strings in the directory.

10. The apparatus according to claim 9, wherein the character codes that are generated by the pattern recognition process are selected from a predetermined set of eight-bit codes, and wherein the respective extension character code comprises a respective eight-bit code that is not included in the predetermined set, and is used by the processor to replace each of the sequences.

11. The apparatus according to claim 9, wherein the pattern recognition process comprises an optical character recognition (OCR) process.

12. The apparatus according to claim 9, wherein the processor is arranged to determine the output string by finding an approximate match between the modified string and one of the known strings, and to output the one of the known strings.

13. The apparatus according to claim 12, wherein the processor is arranged to find the approximate match by computing respective edit distances between the modified string and a plurality of the known strings, and selecting the one of the known strings responsively to the respective edit distances.

14. The apparatus according to claim 13, wherein the processor is arranged to determine respective costs of edit operations involving the extension character code, and to apply the respective costs in computing the respective edit distances.

15. The apparatus according to claim 14, wherein each of the one or more sequences of the character codes is generated due to incorrect segmentation of a respective original character having a respective original character code, and wherein a cost of zero is assigned to a transformation of the respective extension character code associated with each of the sequences to the respective original character code.

16. The apparatus according to claim 12, wherein the directory contains aliases that are derived by replacing each of the one or more sequences of the character codes in the known strings with the respective extension character code, and wherein the processor is arranged to find the approximate match between the modified string

and one of the aliases, and to output the one of the known strings from which the one of the aliases is respectively derived.

17. A computer software product for encoding characters appearing in an area of an image to generate a corresponding output string of character codes, the product comprising a computer-readable medium in which program instructions are stored, which instructions, when read by a computer, cause the computer to receive an identification of one or more sequences of the character codes that are likely to be generated due a segmentation error in application of a pattern recognition process, and to associate a respective extension character code with each of the sequences, and further cause the computer to divide the area of the image into segments such that each segment contains approximately one character, to apply the pattern recognition process to each of the segments in order to generate an input string of character codes, the input string comprising a respective character code for each of the segments, to locate at least one of the sequences of the character codes in the input string, and to replace the at least one of the sequences with the respective extension character code so as to generate a modified string, and to determine the output string by comparing the modified string to a directory of known strings.

18. The product according to claim 17, wherein the character codes that are generated by the pattern

recognition process are selected from a predetermined set of eight-bit codes, and wherein the instructions cause the computer to assign a respective eight-bit code that is not included in the predetermined set to replace each of the sequences.

19. The product according to claim 17, wherein the pattern recognition process comprises an optical character recognition (OCR) process.

20. The product according to claim 17, wherein the instructions cause the computer to determine the output string by finding an approximate match between the modified string and one of the known strings, and to output the one of the known strings.

21. The product according to claim 20, wherein the instructions cause the computer to find the approximate match by computing respective edit distances between the modified string and a plurality of the known strings, and selecting the one of the known strings responsively to the respective edit distances.

22. The product according to claim 21, wherein the instructions cause the computer to determine respective costs of edit operations involving the extension character code, and to apply the respective costs in computing the respective edit distances.

23. The product according to claim 22, wherein each of the one or more sequences of the character codes is generated due to incorrect segmentation of a respective original character having a respective original character code, and wherein the instructions cause the computer to assign a cost of zero to a transformation of the respective extension character code associated with each of the sequences to the respective original character code.

24. The product according to claim 20, wherein the directory contains aliases that are derived by replacing each of the one or more sequences of the character codes in the known strings with the respective extension character code, and wherein the instructions cause the computer to find the approximate match between the modified string and one of the aliases, and to output the one of the known strings from which the one of the aliases is respectively derived.